Bases and spaces

Tutorial

Bases and spaces: resources on the web for accessing the draft human genome - II - After publication of the draft

Colin AM Semple

Address: Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Molecular Medicine Centre, Western General Hospital, Edinburgh EH4 2XU, UK. E-mail: Colin.Semple@ed.ac.uk

Published: 5 June 2001

Genome ***Biology*** 2001, **2(6)**:reviews2001.1–2001.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/6/reviews/2001

Abstract

The volume of human genome sequence and the variety of web-based tools to access it continue to grow at an impressive rate, but a working knowledge of certain key resources can be sufficient to get the most from your genome. This article provides an update to *Genome Biology* 2000, **1**(4):reviews2001.1-2001.5.

The twin publications of the draft human genome sequence [1,2] were accompanied by a large volume of additional information. We were informed that the sequence was comparable to the Apollo moon landings, the invention of the wheel and the Gutenberg printing press. We watched as previously sober members of the media scrambled to calculate the number of telephone directories necessary to hold three billion letters. We sighed as we read another reference to 'designer babies' or scientists 'playing God'. The finer details were invariably lost in this maelstrom. In June 2000, about 24% of the entire genome was available as finished sequence, with only a small proportion finished for most of the chromosomes. Now, in May 2001, about 45% of the genome is in finished form but virtually all of the remainder of human euchromatic DNA is expected to be present in 'draft' form. This increase in sequence has also seen the development of a number of useful web resources that make using the sequence easier.

Putting Humpty back together: genomic sequence assembly

Fragmented genomic sequence is a valuable resource for those interested in a particular gene, but many researchers are interested in a wider region of the genome. For instance, in positional cloning projects it is desirable to know the order and relative orientation of genes, markers and repeats within a given interval. This information can only come from an assembled consensus sequence encompassing the whole region with no gaps, or at least only small gaps of known position and size. Unfortunately this kind of information is available for less than half of the genome; the rest of the sequence exists in unfinished 'draft' form. In practice, this consists of GenBank [3] sequence entries for bacterial artificial chromosome (BAC) clones made up of a number of non-overlapping, arbitrarily ordered, fragments of sequence.

There are two publicly accessible efforts to assemble overlapping fragments from different BAC sequences: the Human Genome Project Working Draft [4] (also known as the 'Golden Path' assemblies) at University of California, Santa Cruz (UCSC), and the National Center for Biotechnology Information (NCBI) Contig Assemblies [5]. The UCSC strategy begins with human genomic sequences from GenBank at a given point (a 'freeze' dataset), ordered and oriented according to the appropriate fingerprinting contigs from Washington University Genome Sequencing Center (WUGSC) [6]. Within each WUGSC contig, draft sequence fragments are assembled into consensus 'raft' sequences using overlaps (detected by the 'ooGreedy' algorithm) between fragments and bridging mRNA, expressed sequence tag (EST), plasmid and BAC-end-pair sequences. Repeated tracts of the letter 'N' are inserted between non-overlapping rafts to give a longer consensus sequence for each WUGSC contig. The NCBI approach also begins by finding an order for adjacent BACs but in this case it is derived from BAC

sequence overlaps (using a variant of the BLAST [7] algorithm), chromosome assignment by fluorescence *in situ* hybridization (FISH) and sequence-tagged site (STS) content. The sequence fragments from these overlapping BACs are then merged into consensus 'meld' sequences. As with the UCSC method, these melds are then ordered and orientated using ESTs, mRNAs and paired plasmid reads, before being combined into a single NCBI genomic sequence contig with melds separated by runs of the letter 'N'. Because NCBI will often be dealing with fewer sequence fragments than UCSC in the construction of a given contig (since NCBI only assemble contigs from overlapping BAC sequences, not from all BAC sequences within an FPC contig (see below) as does UCSC), there should be less opportunity for misassembly (erroneously ordering or orientating sequence fragments).

Celera Genomics [8] has also published a draft assembly for the genome [2]. It is available, under a variety of restrictions, only from their web site. They used a whole-genome shotgun sequencing approach to obtain fragmentary and unmapped sequence data, and combined these with International Human Genome Sequencing Consortium (IHGSC) sequence and mapping data. Celera assembled its own data with those produced by the IHGSC using two different assembly protocols: either with or without mapping information. They found that using the public mapping data gave the assembly greater sequence coverage. The Celera sequences may provide a useful resource for plugging gaps in the IHGSC draft genome; Aach *et al.* [9] found that up to 0.14% of Celera sequence was not present in the public data set. In practice, it can be difficult to retrieve data from the Celera site, presumably as a result of heavy traffic, and BLAST searches of the Celera database can yield a range of slightly odd error messages (my current favorite is the disarming "Your query was not completed because the job required more CPU time than was allotted for the job.").

All assemblies of draft sequences should be treated with suspicion. The UCSC effort [4] is the only one to provide an exhaustive account of the shortcomings of their assemblies. Summary statistics are given for almost every measurable characteristic, including numbers and lengths of gaps and the estimated number of misassemblies. In an elegant series of tests, the UCSC group has examined the performance of their assembly algorithms on artificially produced fragmentary sequence (fragmented sequences from finished regions of the genome). They found that on average their assemblies reproduced the correct ordering of fragments around 85% of the time and the correct orientation of fragments around 90% of the time. But the quality of the assemblies degraded rapidly where the algorithms had to deal with many small fragments, with both correct order and orientation being achieved only 50% of the time (effectively the same as random assembly) in the worst cases. Aach *et al.* [9] also identified possible misassemblies in the Celera assembly. Neither Celera nor NCBI publishes statistics analogous to UCSC's for draft sequence assembly, nor have they published estimates of the rate of error involved.

Bridges over troubled sequence: genomic mapping data

It has been said that the importance of a scientific discovery is proportional to the volume of previous findings it makes redundant. By this measure the ongoing physical mapping effort [6] at WUGSC is extremely important. In combination with others, WUGSC has produced a physical map of the genome using 'fingerprinting' analysis of BACs from a library of genomic sequence clones called RPCI-11 [10]. This map provides the highest resolution human mapping data yet made available and will retain this status until publication of the fully finished human genome sequence. Overlaps between clones are calculated by the program FPC on the basis of clone restriction-fragment patterns, or fingerprints. In this way, fingerprinting-based contigs covering at least 96% of the euchromatic genome have been constructed. The published draft genome sequence assemblies contain hundreds of thousands of gaps, whereas this physical map contains less than 1000. There are therefore many regions in the genome where the physical map can be a useful tool for ordering the chunks of draft sequence we have.

Unless you are dealing with a large region of finished sequence, it is probably wise not to rely on one source of mapping data alone. The International Human Genome Mapping Consortium (IHGMC) physical map [6] can be supplemented with data from other sources. Many fully or partially sequenced BAC clones have been cytogenetically mapped to particular regions using FISH, and this cytogenetic location is given in the GenBank sequence entry for the clone. The end sequences from many BAC clones, including those from libraries other than RPCI-11, are available from The Institute for Genomic Research (TIGR) Human BAC Ends site [11]. In some cases two draft sequences can be ordered relative to one another if they both match different end sequences of a common and otherwise unsequenced clone. The physical and genetic marker content of a sequence of interest can be determined online using the electronic-PCR (e-PCR) [12] program at NCBI. This is a rapid sequence-search algorithm that searches your sequence for occurrences of marker sequences in GenBank. If you enter a BAC clone accession number, the output consists of an ordered list of markers (and the chromosome they map to) down the clone sequence. Further sources of mapping data on the web were described in a previous review [13].

Writing the parts list: genomic sequence annotation

"We've called the human genome the blueprint, the Holy Grail, all sorts of things. It's a parts list," said Eric Lander at the Millennium Evening at the White House, 14 October 1999.

In my opinion this is one of the most accurate public statements about the draft human genome, or at least it will be - once we have reliable genomic sequence annotation. At the moment, we are still rather far from an accurate and comprehensive list of human genes. About half of the human genes identified so far in the IHGSC draft genome [1] are computational predictions supported to a greater or lesser degree by homology to expressed sequences. The IHGSC process of predicting genes is estimated to have a sensitivity of 68-85% (so that 15-32% of genes present in the genome were missed) and an accuracy of around 79% on average (so that 21% of the sequence of each predicted gene was missed) [1].

The valuable work of screening cDNA libraries to verify or refute gene predictions made for the first completed human chromosome, chromosome 22, continues at the Sanger Centre. Shortcuts to the structures of many genes may come from a large collection of full-length mouse cDNAs [14] and large human cDNA collections [15], which are expected to grow rapidly over the next few years. In the meantime the parts list, like the genome sequence itself, will be incomplete. The IHGSC has assembled an initial, non-redundant set of predicted and known human genes called the integrated gene index (IGI), which is to be updated as more data accumulate and is made accessible via the Ensembl site [16]. Beyond the question of where exactly the genes are, it would also be useful to have some idea of what they do. Only around 40% of the IGI-predicted proteins appear to contain known domains - implying that we may be able to predict at least some of their functions - and even in these cases the computational predictions must be supported by lab work before they are regarded as accurate. So, for some years to come many of the elements in the parts list will be unlabeled.

There are now three well-designed websites offering users the chance to browse annotations of the draft human genome. Essentially, all three sites offer a graphical interface to display the results of various analyses, such as gene predictions and similarity searches, for draft and finished genomic sequence. These interfaces are indispensable for allowing rapid, intuitive comparisons between the features predicted by different programs. For instance, one can see at once where an exon prediction overlaps with interspersed repeats or a single-nucleotide polymorphism (SNP). But the three web browsers are not equivalent and there are important distinctions between them, both in terms of the data analyzed and the analyses carried out, as summarized below.

Ensembl

The Ensembl database (Figure 1) [16] was the first to provide a window on the draft genome and started by curating 'confirmed' genes that are computationally predicted (by the Genscan [17] and Genewise [18] gene-prediction programs) and also supported by a significant match to one or more

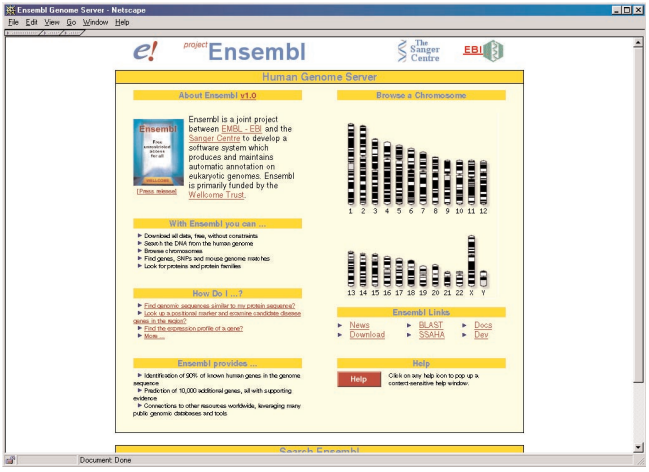

Figure 1
The Ensembl website.

expressed sequences. Ensembl also uses gene structures from public sequence database entries (many of which are experimentally verified), so the total set of Ensembl genes should be a more accurate reflection of reality than computational predictions alone. Many other features have been included as Ensembl has developed: repetitive sequence, cytological bands, genetic markers, CpG island predictions, tRNA gene predictions, UniGene [19] (expressed sequence) clusters, SNPs from the dbSNP database [20], disease genes found in the draft genome (identified by the Online Mendelian Inheritance in Man database, OMIM [21]), and regions of homology to mouse draft genomic sequences. Data retrieval is well catered for, with text searches of all Ensembl entries, BLAST searches of all sequences archived and the availability of bulk downloads of Ensembl data and even software. Ensembl uses the UCSC draft sequence assemblies as its starting point, so its description of a region can only be as accurate as those assemblies allow. Gaps and misassemblies in the genomic sequence could lead to Ensembl missing or wrongly positioning genes and other features. As discussed above, misassemblies are expected to arise most frequently among small BAC fragments, but more broadly the BAC clones mapped to a given region by UCSC (on the basis of FPC data, FISH data, STS content and sequence overlap) should be very accurate. Thus, Ensembl annotation may not be accurate on the finest, local scales but should give a very good idea of what is present in a larger region, perhaps at the level of megabases.

UCSC Human Genome Browser

The UCSC Human Genome Browser (HGB; Figure 2) [4] bears many similarities to Ensembl: it too provides annotation of the UCSC assemblies (and so the same caveats apply) and it displays a similar array of features. There are, however, some additional features of HGB that are not yet found in Ensembl. For example, HGB includes predictions

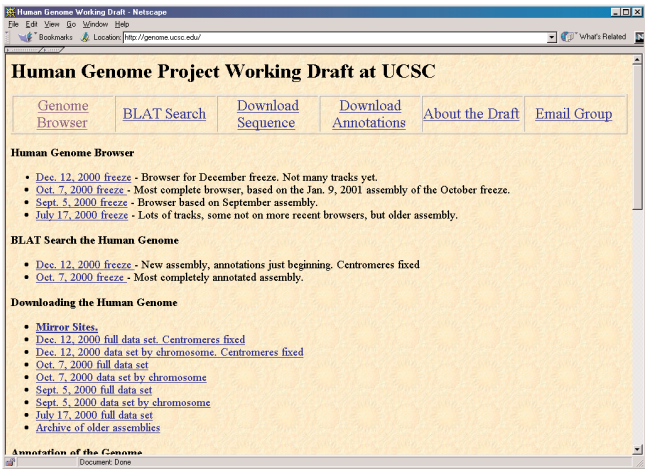

Figure 2
The UCSC Human Genome Browser (the 'Golden Path').

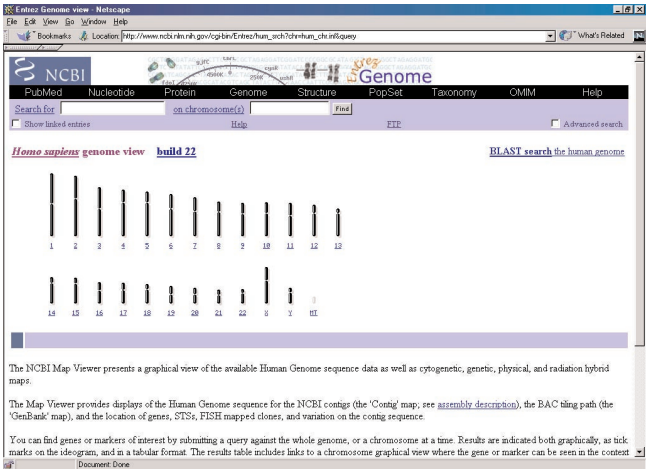

Figure 3
The NCBI Map Viewer.

from more than one *ab initio* gene-prediction program (programs that predict coding sequence on the basis of statistical measures of features such as codon usage, initiation or polyadenylation signals, rather than by homology to known genes) and indicates regions with significant homology to the incomplete genome of the pufferfish *Tetraodon nigroviridis*. These features can provide useful information when dealing with gene predictions that are not well supported by similarity to known mRNA sequences. Another useful feature of HGB is the detailed description of the genomic sequence assemblies. Graphical representations of the fragments making up a region of draft genome can be displayed, showing the relative size and sequence quality of each fragment and also whether any gaps between fragments are bridged by mRNAs or paired BAC end sequences. This means one can get an idea of the likely degree of misassembly in a region. Data retrieval is possible via text, BLAT searches (a faster, less accurate algorithm than BLAST) and bulk downloads of annotation or sequence data.

NCBI Map Viewer

Whereas Ensembl and HGB both show annotation of the UCSC draft genome assemblies the NCBI Map Viewer (NMV; Figure 3) [22] displays features present in the NCBI assemblies. The NMV shows useful comparisons between cytogenetic, genetic and radiation hybrid maps in parallel with NCBI draft and finished sequence contigs. The locations of genes, STSs, and SNPs are indicated on the contig sequences. The NCBI approach to gene prediction is more conservative than those in Ensembl and HGB. No *ab initio* gene-prediction-program is used; instead of known genes, mRNAs and ESTs are aligned to genomic sequence using a program called Acembly. The program also attempts to give alternative splice variants of genes where its alignments suggest them. All annotated genes are connected to NCBI LocusLink [23], which provides links to associated information such as related sequence accession numbers, expression data, known phenotypes, polymorphisms, and so on.

In spite of difficulties with the quality of genomic sequence assemblies, the three browsers discussed above remain extremely useful tools for the cautious biologist. They undoubtedly indicate the presence of most of the coding sequence in a given fragment of genomic sequence and indicate its location in the genome as determined by the best available mapping data. In addition, they have a stab at predicting gene structures for novel genes that should be accurate if the gene in question has a close homolog that is known. Most aspects of the analysis carried out are the subjects of active research, and improvements in performance resulting from the inclusion of new sequence data and algorithms will be ongoing. The downside of these developments is that all annotation of draft sequence is potentially in flux, and one should not assume that the representation of a region will remain the same between different releases of software or data.

Never mind the basics: putting the data to work

As mentioned above, as long as draft genome assemblies and annotation are regarded with a degree of suspicion they will provide hours of entertainment. The necessary suspicion may be formalized as follows. First, always find corroborating data for the position of an annotated feature (for example, make sure a gene's position in an assembly is consistent with any independently derived mapping data in the literature or in OMIM [21] or UniGene [19]). Second, be particularly suspicious of assemblies in regions that are rich in repetitive sequence. Third, bear in mind that duplications (some of which are very recent and extend for tens of kilobases) and pseudogenes are common around the genome. Fourth, never assume that because you retrieve genomic

sequence from an online database (for example the flanking sequence around a SNP) it will not contain repeats. Fifth, never assume that two different websites are using the same versions of draft sequences. Finally, never assume that different releases of data from the same site will be similar or indeed related at all. In short, never take a web server's word for it! Below are some generic approaches to two common tasks people have asked me about.

Characterizing genomic region X

The approach to this task depends on the particular region of interest. If the region is of modest size (say a megabase or less) then you could begin by looking for an NCBI finished or unfinished sequence contig [5] that includes it. If you have sequence from the region, then use the human genome BLAST search at the NCBI [7], and if not, the NMV browser [22] can be used to identify the correct contig according to STS and known-gene content. If you have any doubts about an unfinished contig, examine the BAC sequence accession numbers used to create it and check their positions according to the IHGMC physical map [6,10]. Using the NMV browser you can view known genes and SNPs, but richer annotation is accessible using the UCSC HGB [4] or Ensembl [16] browsers. It is straightforward to identify the corresponding HGB and Ensembl region by searching with the names or accessions numbers of genes or markers within the NCBI contig. By collating the annotations from HGB and Ensembl you will have the most comprehensive publicly available view of the region. Typically, most interest is given to the protein-coding potential of the region.

Where a gene prediction is based on a full-length cDNA clone, there should be little room for error: prediction is reduced to the production of an alignment between the cDNA and genomic sequences. Even in these cases it is possible to end up with an incomplete gene structure, however, if the underlying genomic sequence assembly is gapped or misassembled. It may be helpful to compare the gene structures given by NMV [22] and the other two browsers [4,16], since the NCBI and UCSC genome assemblies presumably contain different gaps and misassemblies. The most unreliable gene predictions are those based on an *ab initio* prediction (for example by Genscan [17] in Ensembl [16]) supported by similarity to transcribed sequence. Unfortunately, although *ab initio* prediction programs are good at detecting the presence of a given gene, they do rather badly at accurately predicting its structure - real exons can be missed and spurious ones can be added. The UCSC HGB browser [4] offers a potential way to resolve problems with spurious additional exons in the form of the output from a second *ab initio* prediction program (Fgenesh). Given that different prediction algorithms produce different falsely predicted exons, you can improve the accuracy of a predicted gene by comparing the results of two algorithms and keeping only those exons predicted by both. In HGB [4] the output of the Fgenesh gene-prediction algorithm can be compared with Ensembl-predicted genes (usually based upon Genscan [17], as discussed above). An additional problem is that some genes can be artificially fragmented into more than one smaller gene by prediction algorithms. In Ensembl [16] this appears as two or more neighboring genes all sharing high similarity to the same sequences according to Ensembl 'supporting evidence' information for the genes.

Automatic annotation systems such as those presented in HGB [4] and Ensembl [16] can also have difficulties differentiating between a functional gene and a recent pseudogene. A recent pseudogene may be sufficiently intact to generate a convincing score within an *ab initio* algorithm and, by definition, will have strong similarity to at least one real gene in the database. Such pseudogenes may carry premature stop codons and appear as truncated versions of real genes - users of HGB [4] and Ensembl [16] beware. Both HGB and Ensembl give an indication of whether a predicted gene appears to be spliced or not by aligning mRNA or EST sequences included in the gene to genomic sequence. The presence of splicing is often a good indication that a predicted gene is functional. Once you have identified a predicted gene that appears convincing, it is a good idea to examine the sequences to which it is similar. Unfinished human genomic sequence can contain contaminating sequence from other organisms: for instance, if a predicted gene resembles a bacterial gene more closely than any vertebrate genes it is almost certainly within contaminated genomic sequence.

At the moment, the best indication of the presence of promoters or other non-coding regulatory elements is conservation of discrete islands of non-coding sequence between human and mouse sequences. Both Ensembl [16] and HGB [4] have incorporated such data as they emerge from the mouse-sequencing project, but only Ensembl provides downloadable mouse sequence data. At the moment, only HGB offers comparisons with a second organism: the pufferfish *Tetraodon nigroviridis*. Both Ensembl and the NCBI intend to make the annotated genomes of other organisms available via their browsers. In the meantime, one can browse genomic regions of mouse-human homology at a low resolution using the NCBI Human-Mouse Homology Map [24]. Alternatively, a more detailed view of the pattern of homology over a region can be obtained using the ingenious PipMaker program [25]. PipMaker aligns two sequences of DNA up to 2 Mb in length and produces a 'percent identity plot' (hence 'pip') showing, at a glance, the pattern of conservation over the region.

All three of the annotation browsers display SNPs within a region of interest, but Ensembl and HGB also show the positions of repetitive sequence, so it is possible to avoid SNPs within repeats for use in PCR-based assays. Unfortunately,

there is no classification of SNPs on this basis, so the user has to plough through all SNPs, making sure the SNP does not overlap with the coordinates given for repetitive sequences. It is also worth remembering that the number of known SNPs is increasing so rapidly that more may have been deposited in dbSNP [20] and/or on The SNP Consortium (TSC) site [26] since the version of the annotation you are browsing was produced. It may therefore be worth searching dbSNP and TSC data with BAC or gene-sequence accession numbers from your region.

Finding all genes coding for members of protein family Y

For a really comprehensive view of all detectable members of a protein family, particularly if it is novel, there is still no substitute for iterative similarity searches and manual examination of alignments. An excellent web-based tutorial outlining this approach is available from William R. Pearson at the University of Virginia [27]. NCBI also offers relevant information on its Educational web pages [28]. For those with more modest aims, such as identifying all human members of a well-studied protein family, the available web resources can remove much of the effort involved. Ensembl [16] is especially useful here, providing a list of the known protein domains in each Ensembl gene derived from the InterPro database [29]. InterPro groups all known proteins on the basis of shared domains and functional sites, and it often includes distantly related members of protein families that it would otherwise be time-consuming to identify. It integrates data from a variety of databases describing protein motifs, domains and families and is the most comprehensive view of relatedness within the human proteome. Thus, the association of Ensembl genes with the appropriate InterPro domains provides a potent means of identifying all genes coding for members of a known protein family. A simple search of the Ensembl database with an InterPro accession number provides a list of Ensembl genes with their locations in the genome - replacing multiple similarity searches, alignment gazing and trawling through disparate sources of mapping data. Of course, all the caveats relating to gene predictions (and proteins conceptually translated from them) discussed above apply. Tales have already emerged that describe problems in Ensembl protein families caused by fragmentation of genes by *ab initio* predictions and the inclusion of contaminating sequence in genomic sequence assembly [30]. The NCBI resources for finding genes coding for protein family members are, at present, rather slender by comparison with Ensembl. A new BLAST page ('BLAST the Human genome' [31]) has been established which allows the user to find the position of a query sequence and its close homologs within the NCBI genomic sequence assembly. The associated information on conserved domains is drawn from the NCBI Conserved Domain Database (CDD). Like the InterPro database, the CDD is built by collating protein family data from other smaller databases, but fewer are included.

References

1. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The Sequence of the Human Genome**. *Science* 2001, **291**:1304-1351.
3. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
4. **Human Genome Project Working Draft** [http://genome.ucsc.edu/]
5. **NCBI Contig Assemblies and Annotation** [http://www.ncbi.nlm.nih.gov/genome/guide/build.html]
6. **Washington University Genome Sequencing Center** [http://genome.wustl.edu/gsc/human/Mapping/]
7. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST/]
8. **Celera Genomics** [http://www.celera.com/]
9. Aach J, Bulyk ML, Church GM, Comander J, Derti A, Shendure J: **Computational comparison of two draft sequences of the human genome**. *Nature* 2001, **409**:856-859.
10. International Human Genome Mapping Consortium: **A physical map of the human genome**. *Nature* 2001, **409**:934-941.
11. **TIGR Human BAC Ends** [http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html]
12. **NCBI UniSTS Electronic PCR** [http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi]
13. Semple, C: **Bases and spaces: resources on the web for accessing the draft human genome**. *Genome Biology* 2000, **1**(4):reviews2001.1-2001.5.
14. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, *et al.*: **Functional annotation of a full-length mouse cDNA collection**. *Nature* 2001, **409**:685-690.
15. Abbott A: **Free access to cDNA provides impetus for gene function work**. *Nature* 2001, **410**:289-290.
16. **Ensembl** [http://www.ensembl.org/]
17. **GenScan** [http://genes.mit.edu/GENSCAN.html]
18. **Genewise** [http://www.sanger.ac.uk/Software/Wise2/]
19. **UniGene** [http://www.ncbi.nlm.nih.gov/UniGene/]
20. **dbSNP** [http://www.ncbi.nlm.nih.gov/SNP/index]
21. **OMIM** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM]
22. **NCBI Map Viewer** [http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch?chr=hum_chr.inf&query]
23. **LocusLink** [http://www.ncbi.nlm.nih.gov/LocusLink]
24. **NCBI Human-Mouse Homology Map** [http://www.ncbi.nlm.nih.gov/Homology/]
25. **PipMaker** [http://bio.cse.psu.edu/pipmaker/]
26. **The SNP Consortium Ltd** [http://snp.cshl.org/]
27. **Exploring Distant Protein Sequence Relationships** [http://www.people.virginia.edu/~wrp/prot_talk12-95.html]
28. **NCBI Education** [http://www.ncbi.nlm.nih.gov/Education]
29. **InterPro** [http://www.ebi.ac.uk/interpro/]
30. Birney E, Bateman A, Clamp ME, Hubbard TJ: **Mining the draft human genome**. *Nature* 2001, **409**:827-828.
31. **BLAST the Human genome** [http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html]